



QUASAR: Quantizing Sanitizer against Adversarial Examples

이름: 권민재
 학번: 20190084
 연구지도교수: 김종

2022. 3. 9.

1 연구 목적

최근 인공신경망을 이용한 이미지 분류는 단순히 이론적인 수준을 넘어 사람들이 많이 사용하는 여러 서비스들에 공격적으로 도입되고 있다. 예를 들어, 온라인 사진 앤솔 서비스들이 사진들에 포함된 얼굴을 분류해주는 기능을 소비자에게 제공하는 것을 넘어, 자율주행차가 카메라를 통해 실시간으로 사물을 인식하고 분류하는 데에도 인공 신경망이 이용되고 있다. 이러한 classifier들이 여러 서비스와 연구의 밑바탕으로 활약하고 있는 것과 달리, 최근 연구들은 classifier들을 속일 수 있는 여러 공격 기법들을 보여주고 있다. FSGM (Fast Gradient Signed Method), PGD (Projected Gradient Descent) 등의 공격 방법을 시효로 이미지 분류 인공 신경망에 대해 다양한 공격 방법이 연구 및 제시되었다. 특정 신경망에 대한 adversarial example을 생성하는 공격 방법이 주를 이뤘으며, 이에 대한 방어 방법 또한 지속적으로 소개되고 있다.

Adversarial example에 대한 방어법은 크게 두 가지 정도로 분류해볼 수 있다. 첫번째로, adversarial example 들을 인공 신경망에 직접 학습시키거나, adversarial example을 분류해낼 수 있도록 학습시키는 방법이 존재한다. 하지만 adversarial example을 직접 이용하는 방식에는 분명한 한계점이 존재한다. 종래의 adversarial attack들은 그 각각의 공격 방법에 따라 매우 다양한 형태의 example들을 생성한다. 이들을 모두 수집하여 예외적으로 인공 신경망에 학습시키는 것은 현실적으로 불가능하다고 판단되기 때문에, 효율적인 방어를 해낼 수 없다. 둘째로, 인공 신경망에 입력되는 이미지를 전처리하는 방법이 존재한다. 전처리를 통해 이미지에 임의로 삽입된 정보들을 상쇄 시켜 example의 perturbation을 줄이는 관점이다. 하지만 이미지를 직접 수정하는 만큼 오히려 classifier를 방해할 수 있다는 지적이 나오고 있다. 예를 들어, **SHIELD**는 이미지 전처리를 이용하는 대표적인 방어법이다. SHIELD 는 JPEG compression을 통해 불필요한 정보를 제거하여 방어를 수행하지만, compression 과정이 분류 체계를 방해하지 않도록 기존 신경망을 "vaccinate" 해야만 했다. [2] 즉, 이미지에 대한 전처리 과정이 인공 신경망을 방해할 뿐만 아니라 방어 시스템 도입의 이식성을 저해하고 있는 것이다.

이에 대해, 기존 **MagNet**에 추가적으로 도입할 수 있는 sanitizer인 **QUASAR**, Quantizing Sanitizer를 제안하고자 한다. MagNet은 detector와 reformer로 대표되는 두 가지 형태의 autoencoder를 통해 인공 신경망의 수정 없이 adversarial example을 탐지하고 전처리하는 방어 방법이다. 지금 시점에서 MagNet이 성공적으로 adversarial example에 대해 저항하고 있는지는 의문스럽다. [1] 하지만 인공 신경망과 독립적으로 구축될 수 있는 시스템이라는 MagNet의 속성은 여타 방어 방법들이 마땅히 따라야 할 분명한 장점이다. 인공 신경망의 수정을 요구하는 방어 솔루션은 그 자체로 분명한 한계가 존재하기 때문이다. 이에, MagNet의 이식성은 그대로 유지한 채로 공격에 대한 저항력을 키우기 위해 MagNet에 추가적으로 도입할 수 있는 sanitizer 모듈인 QUASAR를 제안하고자 한다.

2 연구 배경

2.1 기존 공격법

2.1.1 FSGM (Fast Gradient Signed Method)

FSGM은 일반적인 이미지 x 에 대해서 classifier의 decision boundary를 넘을 수 있는 노이즈가 섞인 adversarial example x' 를 생성하는 것을 목표로 한다. 즉, 이것은 모델 파라미터 θ , x 에 대한 label y 에 대해 loss $J(\theta, x, y)$ 를 최대화시키도록 학습시키는 것을 기본적인 골자로 adversarial example을 획득하는 것이다. 이때, 위에서 구한 loss를 곧바로 이용하는 것이 아니라, x 에 대한 gradient인 ∇_x 를 loss에 곱하여 이미지의 각 픽셀이 loss에 기여하는 정도를 adversarial example에 반영할 수 있도록 하였다. 이를 수식으로 나타내면 아래와 같다. [3]

$$x' = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y))$$

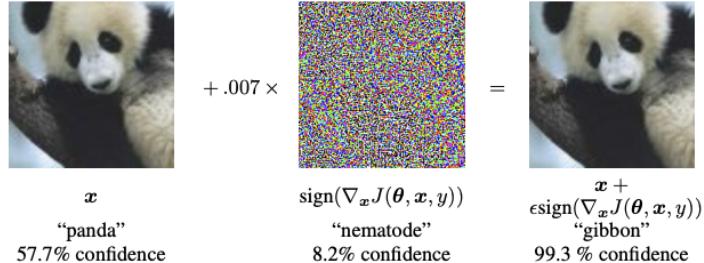


Figure 1: Example of generating adversarial example with FSGM [3]

2.1.2 PGD (Projected Gradient Descent)

PGD는 FSGM과 같이 local maxima를 이용하여 adversarial example을 생성한다. 하지만, 기존 FSGM의 학습 과정을 α 의 크기에 따라 스텝으로 나누고, 그를 반복적으로 수행하여 더 좋은 결과를 이끌 수 있도록 FSGM보다 더 발전하였다. 이러한 PGD를 수식으로 나타내면 아래와 같다. [5]

$$x^{t+1} = \Pi_{x+\mathcal{S}}(x^t + \alpha \text{sign}(\nabla_x L(\theta, x, y)))$$

2.1.3 DeepFool

DeepFool은 앞서 소개한 FSGM이나 FGSM과 상이한 전략을 채택했다.[6] 앞선 방법들이 loss를 최대화 하는 방향으로 perturbation을 생성하는 것과 달리, DeepFool은 입력된 이미지와 가장 가까운 decision boundary를 찾은 후 boundary를 넘어서 classifier가 잘못된 결정을 내리도록 유도한다. 즉, 가장 가까운 boundary에 이미지를 사영하는 방식을 통해 perturbation을 더하는 것이다. 다만, 대부분 인공 신경망의 decision boundary가 다차원의 곡선의 형태이기 때문에, DeepFool은 해당 곡선들을 linear하게 근사하는 과정을 포함하고 있다.

2.2 기존 방어법

2.2.1 MagNet

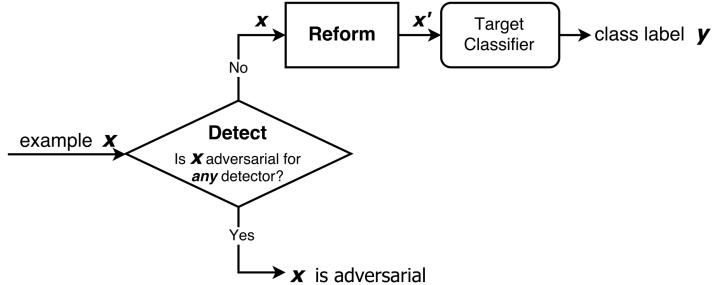


Figure 2: Two autoencoders of MagNet [1]

MagNet은 adversarial example을 이용한 공격을 효과적으로 방어하기 위해서 **Detector**와 **Reformer**를 활용한다.[1] Detector와 reformer는 둘 다 인공 신경망을 학습시킨 데이터를 통해 학습시킨 autoencoder들이다. Detector는 입력된 이미지와 출력된 이미지의 변화된 정도를 통해 해당 이미지가 adversarial example인지 아닌지 확인하는 장치이다. 두 이미지의 차이가 일정 임계 정도를 넘으면 adversarial example이라고 판단하는 방식이다. 그리고 reformer는 입력받은 이미지를 autoencoder를 통해 인공 신경망이 학습한 데이터에 더 가깝게 변환시키는 역할을 한다. 이를 통해 이미지에 섞인 perturbation을 제거하는 것을 목표로 한다. MagNet이 별로 공격에 효과가 없다는 연구[1] 또한 존재하나, 인공 신경망을 직접적으로 수정하지 않고도 방어 체계를 구축할 수 있음을 보였다.

2.2.2 SHIELD

SHIELD는 JPEG 압축을 통한 이미지 전처리를 통해 adversarial example의 perturbation을 제거하는 것을 목표로 하는 방어 체계이다.[2] SHIELD는 공격을 효과적으로 방어하기 위해 SLQ (Stochastic Local Quantization)이라고 이름 붙인 기술을 도입하였다. SLQ는 이미지를 그리드로 나누어서 각 칸을 랜덤의 압축률로 압축한 뒤 다시 하나의 이미지로 합치는 기술이다. 이 기술을 통해 perturbation을 최대한 억제할 수 있도록 만들었다고 연구진은



Figure 3: Stochastic Local Quantization of SHIELD [2]

설명한다. 하지만 SLQ의 전처리 과정은 classifier를 오히려 방해하였고, 이를 해결하기 위해 연구진은 "vaccinate", 즉 기존 인공 신경망에 JPEG로 압축된 이미지를 추가적으로 학습시키는 과정을 거쳐야 했다. 압축을 통해 정보를 줄이는 의도는 좋았으나, SLQ의 압축 과정이 분류 체계를 방해하지 않도록 인공 신경망을 vaccinate해야한다는 점이 이식성을 저해하여 아쉽게 평가된다.

2.2.3 Feature Distillation

Feature Distillation은 기존 JPEG의 압축 과정을 개량하여 vaccinate 과정 없이 전처리할 수 있도록 함을 목표로 한다.[4] JPEG의 압축 과정은 사람의 눈에 보이기에 필요 없는 부분을 제거하는 것이지, 이것이 인공 신경망을 타겟팅했다고는 말하기 힘들다. 연구진은 이 특성을 이용하여, JPEG의 압축 과정에서 쓰이는 DCT와 양자화 과정을 인공 신경망에 맞게 개량하였고, 이를 통해 유의미한 이미지 전처리를 할 수 있었다.

3 연구 방법

QUASAR는 MagNet에서 reformer에 이미지가 입력되기 이전에 perturbation을 최소화할 수 있도록 전처리하는 모듈형 sanitizer를 목표로 한다. 즉, MagNet의 detector와 reformer 사이에 위치하여 이미지를 전처리하는 엔진으로 QUASAR가 사용되는 것이다. QUASAR는 [4]의 방법을 준용하여 인공 신경망을 겨냥한 DCT와 양자화 기법을 이용하여 이미지를 전처리할 예정이다.

QUASAR의 구현과 실험을 위해 타겟이 될 인공 신경망과 위협 모델이 필요하다. 타겟이 될 인공 신경망으로는 MNIST를 분류하는 인공 신경망과 CIFAR-10을 분류하는 인공 신경망을 선택할 것이다. 범용적으로 실험에 많이 이용되는 데이터셋임을 고려하여, 이들을 분류하는 인공 신경망을 공격 대상으로 선택하였다. 그리고, 이들을 공격하는 위협 모델으로는 FSGM, PGD, 그리고 DeepFool의 구현체를 선택할 것이다. 셋의 공격 방법이 Classifier의 공격에 널리 사용됨을 고려하여 셋의 공격에 대해 QUASAR가 얼마나 잘 방어할 수 있을지 측정할 것이다. QUASAR의 방어 수준은 기존 MagNet과의 비교를 통해 평가할 것이다. 앞서 선정한 방어할 두 상황과 세 가지의 위협 모델에 대해 아무런 방어 요소가 없을 때, MagNet을 통해 방어를 할 때, 그리고 MagNet에 QUASAR를 장착하였을 때의 상황을 비교하여 QUASAR의 성능을 평가할 예정이다.

4 기대 효과

이식성이 높은 인공 신경망 방어 체계를 구축하기 위해서는 보호하고자 하는 인공 신경망의 수정을 최소화해야 한다. 즉, 인공 신경망의 방어 체계는 기존 인공 신경망에 모듈형으로 방어 모듈을 추가하는 방향으로 성장해야함을 의미한다. 이러한 관점에서, QUASAR는 MagNet을 더 발전시킬 수 있는 모듈으로써 역할할 뿐만 아니라, 인공 신경망 파이프라인의 적재적소에 추가될 수 있는 모듈형 방어 체계를 QUASAR를 통해 제시하고자 한다. 일차적으로는 MagNet에 QUASAR를 추가하여 MagNet의 방어 성능을 드높이는 것을 목표로 삼지만, 앞으로 더 나아가 이차적으로는 여타 다른 인공 신경망 파이프라인에도 큰 무리 없이 QUASAR가 장착되어 방어 성능을 더 높일 수 있도록 만들 것이다.

5 연구 추진 일정

해당 연구를 수월하게 수행하기 위해서 아래 표와 같은 일정을 수립하였다.

2022. 03. 11.	연구 제안서 제출
2022. 03. 31.	보호하고자 하는 인공 신경망 모델 생성
2022. 04. 11.	타겟 인공 신경망에 대한 위협 모델 구현
2022. 04. 22.	연구 진행 보고서 제출
2022. 05. 11.	QUASAR 구현
2022. 05. 20.	타 방어 기법과 비교 및 평가
2022. 06. 01.	연구 결과 보고서 제출

참고문헌

- [1] Nicholas Carlini and David Wagner. Magnet and” efficient defenses against adversarial attacks” are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017.
- [2] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–204, 2018.
- [3] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.
- [4] Zihao Liu, Qi Liu, Tao Liu, Nuo Xu, Xue Lin, Yanzhi Wang, and Wujie Wen. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 860–868. IEEE, 2019.
- [5] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.
- [6] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016.